AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes

Pedro González-Torres,[a,b,c] Francisco Rodríguez-Mateos,[d] Josefa Antón,[a,e] Toni Gabaldón[b,c,f]

[a]Department of Physiology, Genetics, and Microbiology, University of Alicante, Alicante, Spain
[b]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Barcelona, Spain
[c]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), Barcelona, Spain
[d]Departamento de Matemática Aplicada, University of Alicante, Alicante, Spain
[e]Multidisciplinary Institute of Environmental Studies Ramon Margalef, University of Alicante, Alicante, Spain
[f]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ABSTRACT Homologous recombination (HR) enables the exchange of genetic material between and within species. Recent studies suggest that this process plays a major role in the microevolution of microbial genomes, contributing to core genome homogenization and to the maintenance of cohesive population structures. However, we still have a very poor understanding of the possible adaptive roles of intraspecific HR and of the factors that determine its differential impact across clades and lifestyles. Here we used a unified methodological framework to assess HR in 338 complete genomes from 54 phylogenetically diverse and representative prokaryotic species, encompassing different lifestyles and a broad phylogenetic distribution. Our results indicate that lifestyle and presence of restriction-modification (RM) machineries are among the main factors shaping HR patterns, with symbionts and intracellular pathogens having the lowest HR levels. Similarly, the size of exchanged genomic fragments correlated with the presence of RM and competence machineries. Finally, genes exchanged by HR showed functional enrichments which could be related to adaptations to different environments and ecological strategies. Taken together, our results clarify the factors underlying HR impact and suggest important adaptive roles of genes exchanged through this mechanism. Our results also revealed that the extent of genetic exchange correlated with lifestyle and some genomic features. Moreover, the genes in exchanged regions were enriched for functions that reflected specific adaptations, supporting identification of HR as one of the main evolutionary mechanisms shaping prokaryotic core genomes.

IMPORTANCE Microbial populations exchange genetic material through a process called homologous recombination. Although this process has been studied in particular organisms, we lack an understanding of its differential impact over the genome and across microbes with different life-styles. We used a common analytical framework to assess this process in a representative set of microorganisms. Our results uncovered important trends. First, microbes with different lifestyles are differentially impacted, with endosymbionts and obligate pathogens being those less prone to undergo this process. Second, certain genetic elements such as restriction-modification systems seem to be associated with higher rates of recombination. Most importantly, recombined genomes show the footprints of natural selection in which recombined regions preferentially contain genes that can be related to specific ecological adaptations. Taken together, our results clarify the relative contributions of factors modulating homologous recombination and show evidence for a clear a role of this process in shaping microbial genomes and driving ecological adaptations.

Citation González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. 2019. Impact of homologous recombination on the evolution of prokaryotic core genomes. mBio 10:e02494-18. https://doi.org/10.1128/mBio.02494-18.

Editor Joseph Heitman, Duke University

Address correspondence to Toni Gabaldón, tgabaldon@crg.es.

Received 10 November 2018
Accepted 30 November 2018
Published 22 January 2019

KEYWORDS comparative genomics, core genomes, genome evolution, intraspecific diversity, homologous recombination

Gene repertoires in prokaryotic genomes can be divided into the core genome, comprising genes ubiquitously present in all strains of a species, and the accessory (or flexible) genome, comprising genes whose presence is variable within a species (1, 2). Evolutionary analyses of prokaryotic genomes have revealed that changes in the accessory genome are often associated with horizontal gene transfer (HGT) and site-specific recombination involving mobile genetic elements, whereas changes in the core genome generally involve vertical transmission and homologous recombination (HR) (3, 4). Furthermore, recent studies revealed that prokaryotic populations are structured in the form of cohesive clusters, each composed of strains with genomic similarity levels above a given threshold, which are separated from other clusters by much larger genetic distances (5, 6). This phenomenon has been observed in several taxonomic groups and in diverse environments (5, 7–9). However, what mechanisms generate these clusters and what evolutionary forces drive genetic cohesion in prokaryotic populations are still not fully understood. In this regard, two main hypotheses have been proposed: (i) the neutral model (10), which highlights the role of HR as a passive, neutral mechanism driven solely by genetic divergence, and (ii) the ecotype theory (11, 12), which emphasizes the role of natural selection and/or genetic drift in maintaining groups with similar ecological features. Note that the two models are not mutually exclusive but rather could act synergistically, with the relative contributions of the models differing across environments or clades (5). Besides the issue of what evolutionary forces could create cohesive population clusters, many other issues remain open. For instance, it is as yet unknown whether HR occurs indiscriminately across the whole genome or whether selection shapes HR patterns. In addition, we still have only a very sparse picture of the impact of HR and how this varies across phylogenetic groups or lifestyles. To address such issues, broad analyses that cover diverse clades and lifestyles under the same analytical framework are needed (1, 13–15). However, genomic analyses focused on intraspecific HR processes are scarce due to technical limitations (16, 17). In addition, most studies have been based on multilocus sequence analysis (MLSA) or considered different methodologies (13, 16) or few species or focused on individual genomic factors (18–20). Here we set out to assess intraspecific genomic exchange occurring through HR on a taxonomically broad data set and to perform a comprehensive analysis of a wide range of functional, ecological, and genomic factors. To our knowledge, this was the first such broad study to use a common methodological framework and to consider a broad range of potentially relevant genomic factors.

RESULTS

Quantification of homologous recombination. We retrieved 338 genomes belonging to 54 bacterial and archaeal species—as defined by the 16S-divergence criterion (21, 22)—spanning a wide range of taxa, environments, and lifestyles—as defined by previous studies (13, 16) (see Materials and Methods) (Fig. 1; see also Table S1A and B in the supplemental material). To enable unbiased comparisons, we used a common methodology to reannotate each genome and quantify parameters based on genomic composition, within-species sequence variation, ecological specialization, and other factors for which a role has been proposed or presumed but whose relationship with HR remains poorly assessed (see Materials and Methods and Table S1B). As the main focus of our study was the comparison across clades and lifestyles with the aim of identifying trends in functional, ecological, and genomic factors related to HR, we prioritized a broader taxonomic sampling over a high number of strains per species. Finally, the sampling set was similar in size and taxonomic distribution to that used in recent studies focused on interspecies exchanges (19).

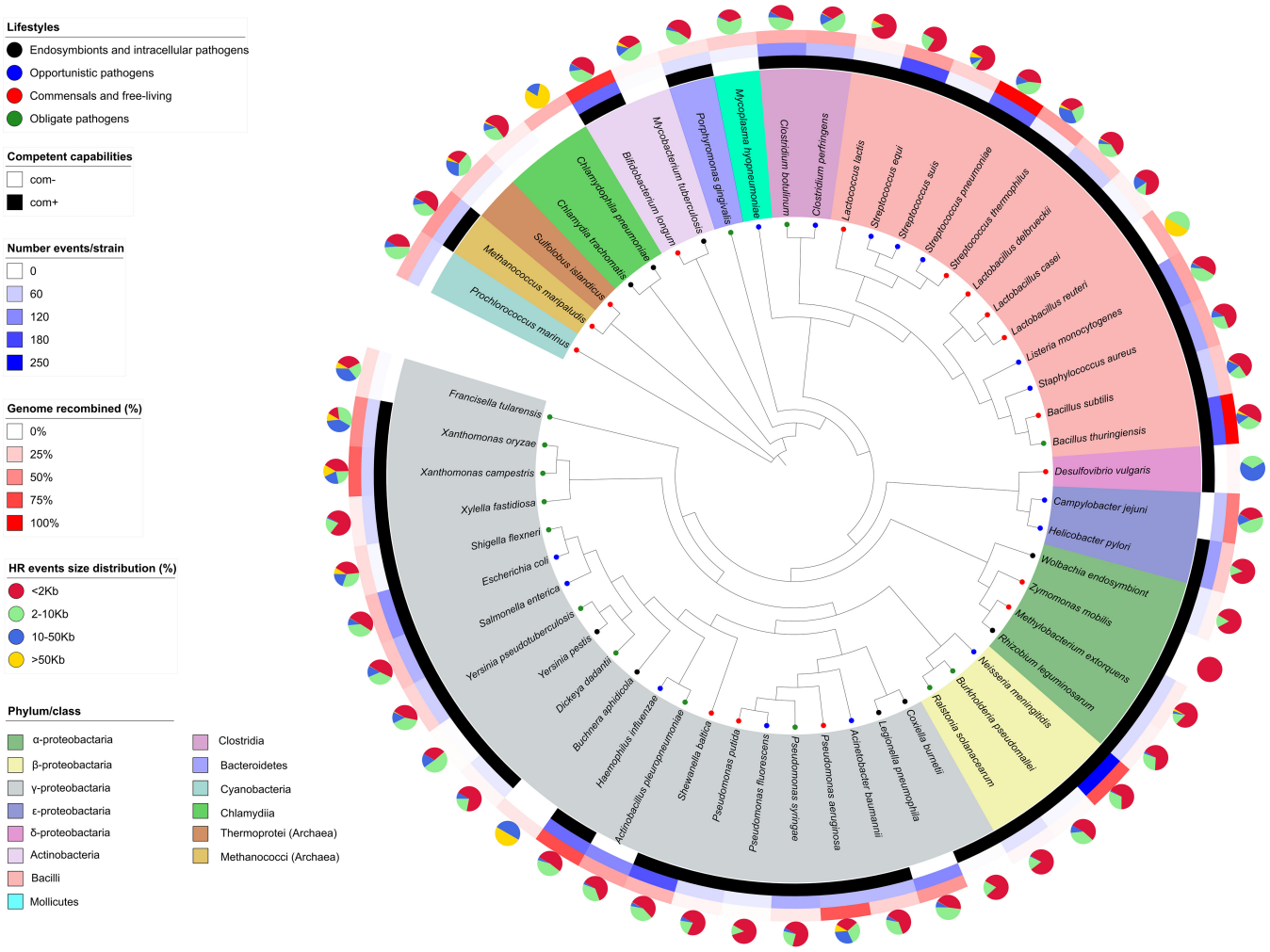

FIG 1 Data set composition. A 16S rRNA phylogenetic tree of the 54 prokaryotic species included in this study is shown. The tree was drawn using itool (https://itol.embl.de/). The innermost circle layer shows the species nape and associated clade. Analysis of HR events was performed. The innermost layer indicates the competence capability. The second (i.e., contiguous) layer, third layer, and fourth layer correspond to the number of HR events per strain, the proportion (%) of recombined genome, and the size distribution (%) of HR events.

We next inferred HR events and related parameters for each species by scanning their aligned genomes with a pipeline that combines several HR detection algorithms and selects the events consistently predicted by several methods (see Materials and Methods and Table S1B). This pipeline comprises state-of-the-art methodologies used in recent analyses (19, 23) of HR in microbial genomes, such as RDP4 v4.15 (24, 25) and ClonalFrame 1.2v (26, 27). In brief, these methodologies use a sliding-window approach for the detection of phylogenetically incongruent regions from genome alignments. Different methods implemented in these programs differ in their specific parameters, algorithms, and tests for significance. Therefore, we opted for a conservative approach that selected only those regions that were positive in at least three of the five methods used. In total, we detected 16,300 HR events that were distributed nonhomogenously across the analyzed species (Fig. 1). We found no significant correlation between the number of strains sampled per species and the number of events detected per strain ($P < 0.05$, $r^2 = 0.048$ [Pearson]) or the percentage of recombined genome ($P < 0.05$, $r^2 = 0.031$ [Pearson]), indicating that differences in sample size did not affect our results regarding differences between species.

With some exceptions, and for the species evaluated in previous studies, our estimates were congruent with those reported based on multilocus sequence analysis (MLSA) (Table S1C) or complete genomes (Table S1D), where differences with MLSA

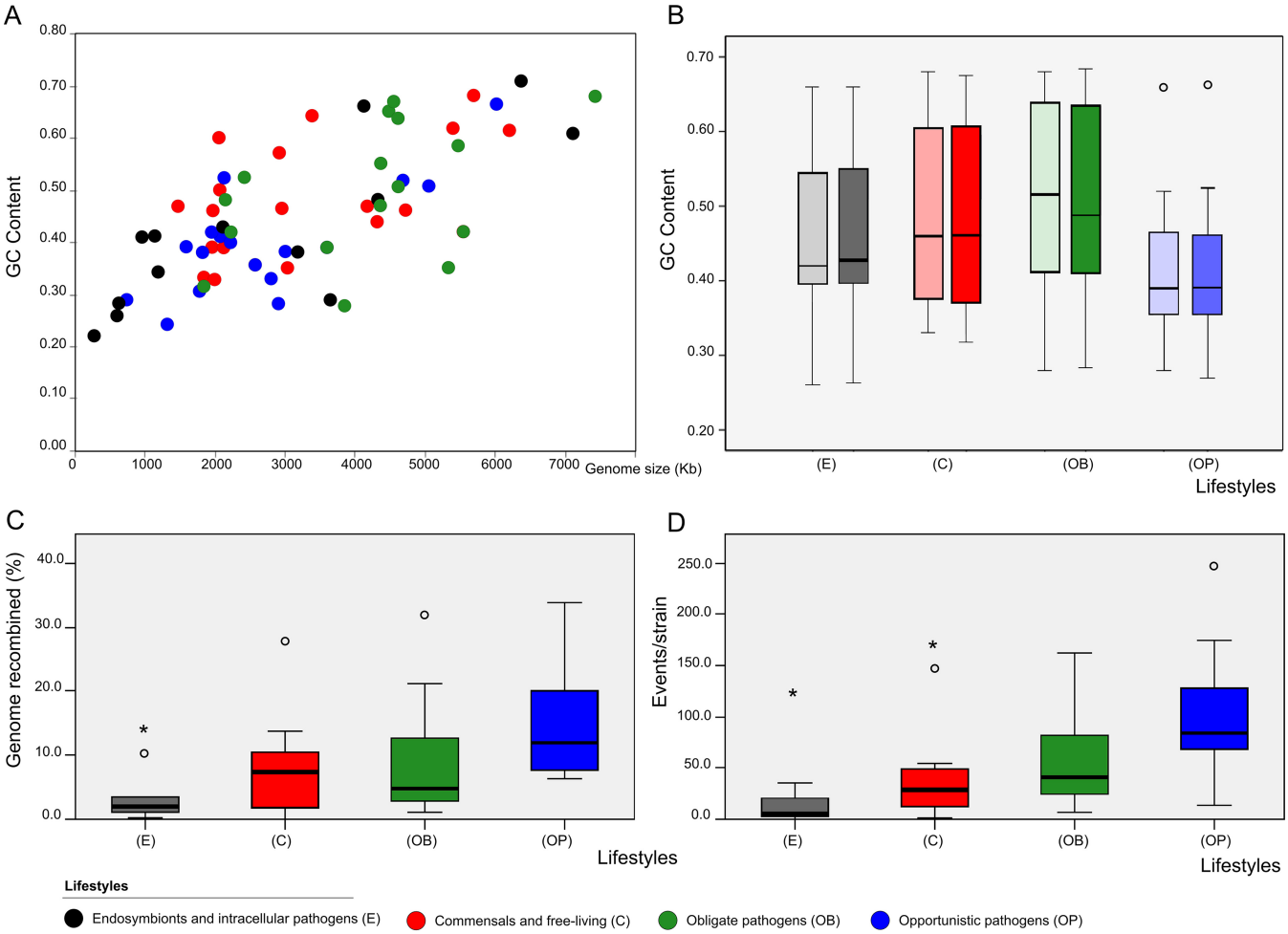

FIG 2 HR characteristics and lifestyle effect. Four lifestyles are represented in all the figures by the same color code: black, endosymbionts and intracellular pathogens; blue, opportunistic pathogens; red, commensal and free-living pathogens; green, obligate pathogens (green). (A) GC content among 54 species included in this study distributed in 4 lifestyles. (B) Box plot comparing average levels of GC content in recombinant events (solid color; right paired boxes) and whole genomes (grayed-out color; left paired boxes). (C and D) HR distribution (events/strain) (C) and proportion of genome recombined (D) based on lifestyle distributions (both $P < 0.05$ [Kruskal-Wallis and Jonkheere-Tepstra tests]).

may be attributable to the low number of genes typically considered in that methodology (13, 28, 29). The detected HR events differed in the length of the recombined region (i.e. size) (Fig. 1; see also Fig. S2 in the supplemental material), which ranged from 0.1 kb to more than 200 kb. Although only 11% (1,799) of the sequences corresponding to the detected events were longer than 10 kb and 4.5% (75) longer than 80 kb (Fig. S1 [additional file 2]), the proportion of events with large sequences was considerable in some species, in accordance with previous studies (Fig. 1; see also Table S1C) (30–32). In contrast to what had been suggested in previous studies (18) and to what is typically observed in eukaryotes, we did not find significant differences in GC content between regions affected by HR events and the rest of the genome (Fig. 2; see also Table S1B).

Influence of phylogeny and lifestyle. The effect of the phylogenetic distance on the incidence of recombination has been widely discussed (3, 13, 16). Most studies suggest an inverse relationship between HR rate and sequence divergence (3, 10, 33). In agreement with this, we observed very low recombination rates in the control groups, which comprised genomes from different species of the same genus (Table S1A and D). Previous studies had also suggested that recombination rates may be similar between closely related species (Table S1C), perhaps due to the existence of evolutionary constraints at the genus level (13, 16). However, these studies included only two

species that shared lifestyle and therefore did not control for this factor. We tested this hypothesis in our data set, focusing on seven genera comprising several species, among which four comprised different lifestyles (Table S1B). We observed large differences in recombination levels within many of the genera considered. An illustrative example is the genus *Yersinia*, which includes the intracellular human pathogen *Y. pestis* and the opportunistic pathogen *Y. pseudotuberculosis*; the latter species showed a much higher recombination/mutation (r/m) rate and 10 times more HR events per strain. Similar trends were observed in the genera *Streptococcus* and *Pseudomonas*, in which HR levels were lower in the free-living species than in their pathogenic relatives. Taken together, our data suggest that phylogenetically closely related prokaryotic species with different lifestyles could present contrasting recombination levels, pointing to an influence of lifestyle in HR patterns.

To gain further insights into the relationship between recombination and lifestyle, we compared HR levels among the members of the four broad lifestyle groups included in our data set (Fig. 2). This lifestyle classification was taken from previous MLSA studies (13, 16), and species were assigned based on the JGI (Joint Genomics Institute) database metadata for each strain. Our results revealed significant differences across lifestyles ($P < 0.05$ [Kruskal-Wallis and Jonkheere-Tepstra tests]). The class containing "endosymbionts and intracellular pathogens" showed the lowest HR levels, followed by "commensals and free-living," and "obligate pathogens" and with "opportunistic pathogens" showed the highest levels. Of note, this relative order corresponds to those reported from studies for interspecific HGT (34, 35), suggesting that similar constraints may act at the intra- and interspecies levels. Despite this trend, HR levels ranged widely within each lifestyle class (Fig. 1 and 2; see also Table S1D).

Genomic variables related to ecological specialization and intraspecific DNA exchange. We next considered different genomic variables that have been found to be related to processes of adaptation and ecological specialization. These include homologous repair systems and competence capabilities as well as defense systems such as toxin-antitoxin, abortive infection (36), restriction modification (RM) (19), and clustered regularly interspaced short palindromic repeat (CRISPR)-Cas (37) systems. We noted that the observed HR trend across lifestyles parallels that observed previously for the content of *rec* genes and the same four lifestyles (38). In this regard, it has been hypothesized that the abundance of Chi sequences (*rec* gene target) should correlate with HR levels (13). We tested this using strains of seven pathogenic species for which Chi patterns were available. Our results confirmed a positive correlation of Chi densities (13) and HR prevalence (Fig. S2). This finding is reinforced by the fact that these species are related only distantly and that their respective Chi sequences are thought to have originated independently (39). This strong coevolution supports the idea of a functional link between *recA* and HR, through the recognition of the Chi-RecBCD complex in Gram-negative bacteria and of Chi-AddAB in Gram-positive bacteria, which act as substrates for the homologous pairing by RecA protein (40, 41).

Besides their essential role in the maintenance of DNA integrity by means of HR, *rec* genes participate in the integration of DNA acquired by conjugation and transformation (3, 42, 43). We explored relationships between genetic acquisition mechanisms and HR levels by classifying species into qualitative competent and noncompetent groups on the basis of metadata available in the Integrative Microbial Genomes (IMG) database and of the presence of the *com* regulon genes (44). Although the fractions of recombined genome were similar in the two groups ($P > 0.05$ [Jonckheere-Tepstra test]), the number of events per strain was significantly larger in competent species, especially when events associated with shorter sequences were considered (Fig. 3) ($P < 0.05$ and $r^2 = 0.317$ for <10-kb linear correlations; $P < 0.01$ and $r^2 = 0.414$ for <2-kb linear correlations). Furthermore, we found significant correlations between the number of short event fragments and the abundance of transposable elements ($P < 0.05$ and $r^2 = 0.72$ [linear correlation]) and between *com* gene content and transposable elements ($p < 0.05$ $r^2 = 0.37$, linear correlation). Taken together, these results point to a role of competence in the exchange of short fragments, while conjugative processes may

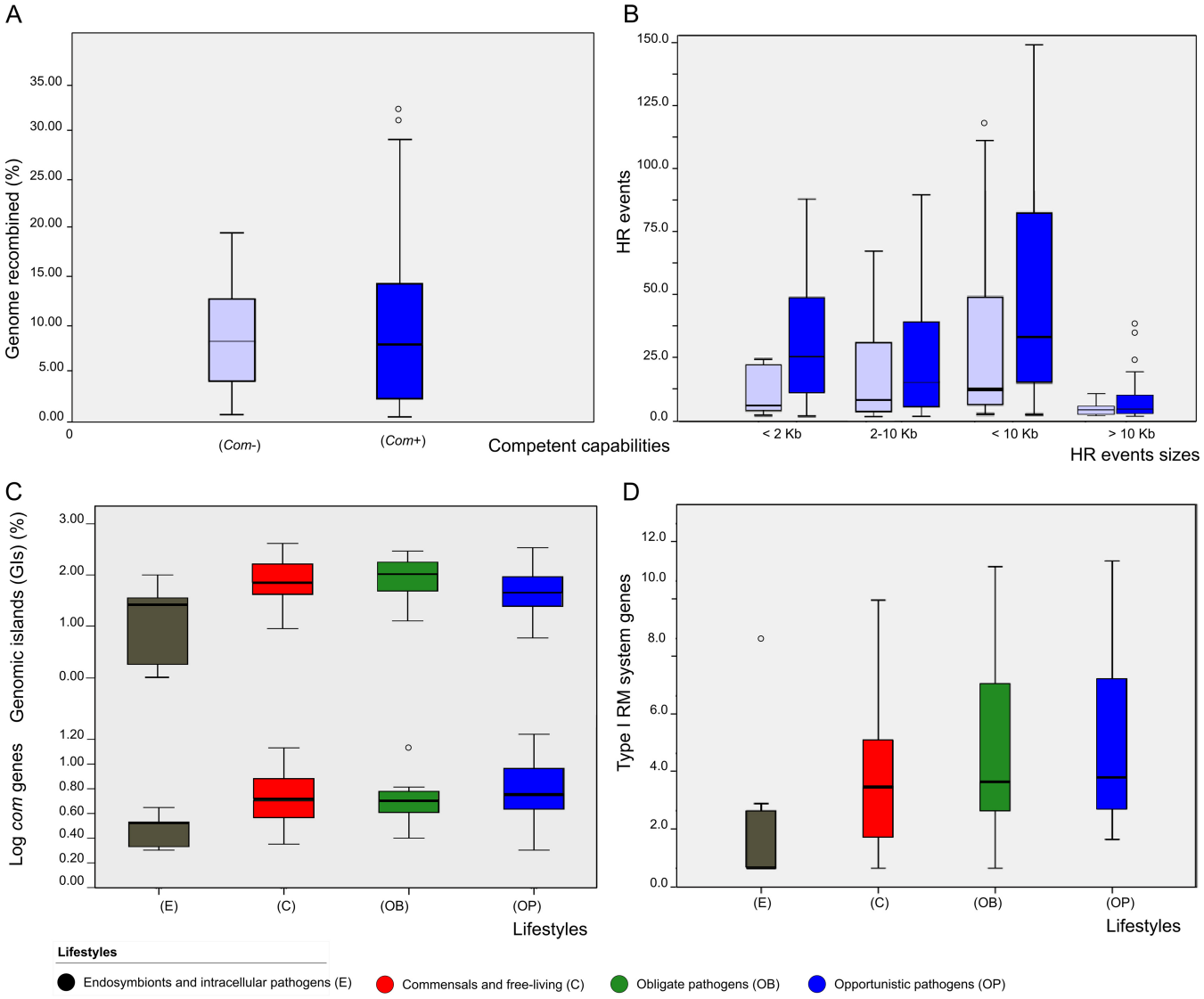

FIG 3 Effect of genomic variables on HR distribution. (A) Proportion (%) of genome recombined based on competence capabilities (*Com*+/*Com*−). (B) HR event fragment size distribution based on competence capabilities: competent (*Com*+) (solid color, right paired boxes) and noncompetent (*Com*−) (grayed out color, left paired boxes). (C and D) Genomic island (GI) distributions (%), *com* gene content (C), and type I restriction modification system (RM-I) gene content distribution (D) based on the different lifestyles considered. Four lifestyles are represented with the following color code: black, endosymbionts and intracellular pathogens; blue, opportunistic pathogens; red, commensal and free-living pathogens; green, obligate pathogens (green).

largely be involved in the exchange of longer (>10-kb) fragments, as previously suggested from isolated cases (3, 16). Finally, we observed a significant positive correlation of the number of short event fragments (2 to 10 kb) with both *com* gene content and genomic islands (GIs) ($P < 0.05$, $r^2 = 0.43$ and $r^2 = 0.40$ [Spearman's rho]), which are considered hot spots favoring both HR and HGT exchanges (45) and which were more abundant in pathogens and free-living species (Fig. 3).

Defense systems such as restriction modification (RM) or CRISPR-Cas systems act as barriers modulating DNA acquisition and recombination fluxes (19). We detected significant differences in RM gene content across lifestyles ($P = 0.001$ to <0.05 [Kruskal-Wallis and Jonkheere-Tepstra tests]), even in comparing different RM systems (type I to type III) individually ($P < 0.05$) (Fig. 3; see also Fig. S3). This trend parallels those reported for *rec* genes and HR events among lifestyles such that opportunistic pathogens and free-living species show higher values. Furthermore, we observed a significant positive correlation between the number of events per strain and the

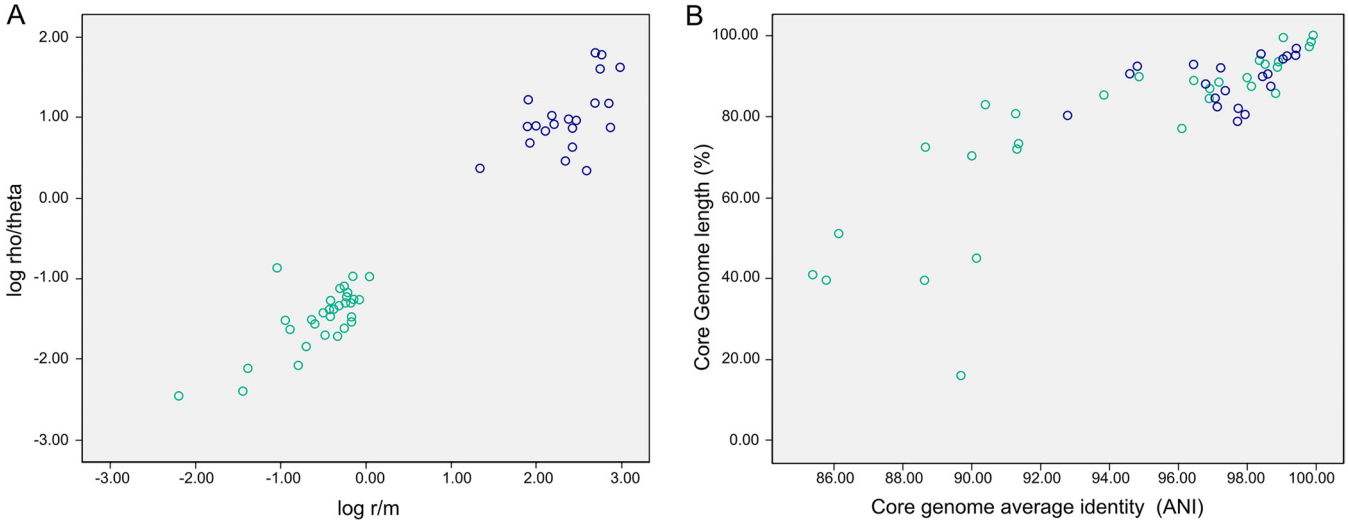

FIG 4 Influence of lifestyle and role of HR in population structure and evolution. (A) Correlation of r/m and rho/theta ratios (r/m of >1 [blue] or <1 [green]) for 54 species and (B) distribution based on core genome length and core genome identity (ANIb).

content of type I and II RM systems ($P < 0.05$, $r^2 = 0.3$ and $r^2 = 0.43$, respectively [Spearman's rho]). Taken together, these results indicate that species encoding more RM systems tend to acquire more genetic material by HR. This observation is compatible with the role of RM systems generating DNA recombinogenic extremes that may promote HR (46), although it has been observed that intraspecific exchanges are limited between strains coding for different RM type systems (47). In this regard, higher diversities of CRISPR-Cas system genes tended to be associated with lower HR levels (Fig. S4). This suggests that intraspecific diversity in CRISPR-Cas systems, which generally act against heterologous sequences, may as well affect homologous sequences exchanged by conjugation or transformation. Of note, a possible form of coevolution between competence genes and CRISPR-Cas systems has been recently suggested in which loss of competence is followed by loss of CRISPR-Cas systems (48).

Role of HR in population structure and evolution. We explored the evolutionary impact of HR in the core genomes by assessing the relationship between the overall genome average nucleotide identity (ANI) based on BLAST (ANIb) and r/m and rho/theta ratios (where r/m is the ratio of probabilities that a given site was altered through recombination [r] and mutation [m] and measures how important the effect of recombination—relative to mutation—was in the diversification of the sample and rho/theta is the ratio of rates at which recombination [rho] and mutation [theta] occurred and is a measure of how often recombination events happen relative to mutations) (Fig. 4). Most species with positive logarithmic r/m values (r/m >0.33; blue dots in Fig. 4) showed ANIb values greater than 95% (Fig. 4), below which there was a significant fall in ANIb values and in the proportion of synthetic core genome regions. This value corresponds to the threshold observed in metagenomic studies for species with cohesive population structures (49–51). Comparisons between strains of the same species have revealed the same clusters and differences at ANIb values between 1% and 5% (5, 6), corresponding to a dominant evolutionary role for HR. In addition, 95% identity corresponds to a DNA-DNA hybridization value of 70%, above which two strains are considered to be from the same species and below which there is a sharp decrease in HR efficiency (14, 21, 52, 53). The average ratio of nonsynonymous substitutions to synonymous substitutions (dN/dS) across different lifestyles (Fig. S5 [additional file 2]) revealed that opportunistic pathogens and commensal species with large effective populations and high levels of HR, such as *Escherichia coli* (30), showed lower dN/dS values than obligate pathogens such as *Chlamydia pneumoniae*, even among species of the same genus. Although earlier observations of particular species

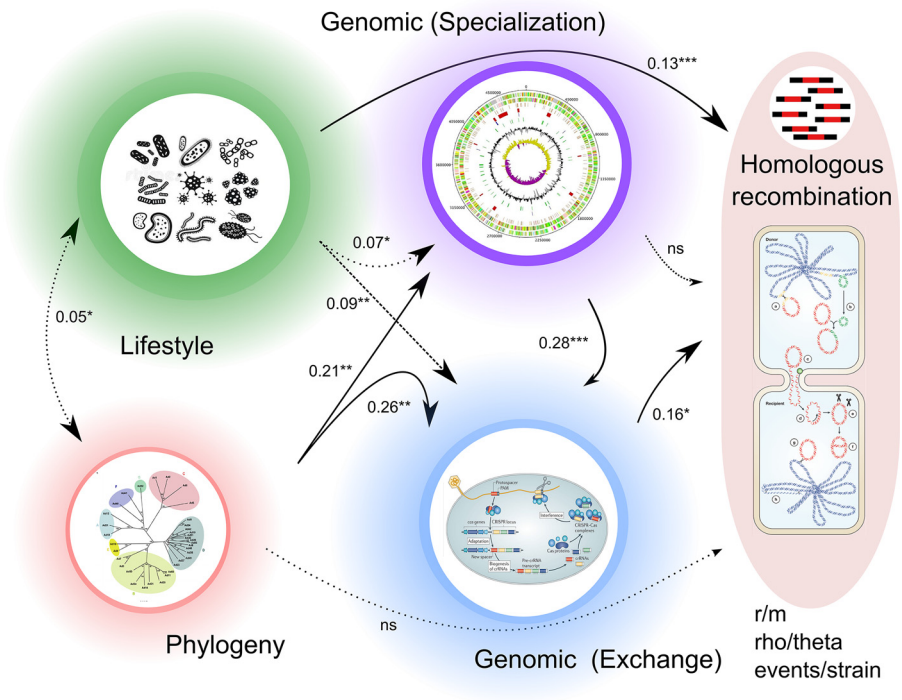

FIG 5 General model. A scheme is presented of a path analysis model proposed for the analysis of the influence of lifestyles, phylogeny, barrier/motility genomics variables, and genomics characteristics on HR levels detected among 54 species analyzed. Significant *r* values (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$) for partial comparisons carried out during Mantel test are indicated. Arrows indicate relationships between variables, with the thickness of the arrows being proportional to the correlation between the connected variables.

(54–56) suggested this relationship, here we uncovered a general trend supporting a role of HR as the main evolutionary force—among other variables considered in this study. This important role of HR is predicted by the neutral model (10).

General model. We used general linear regression and path analysis to integrate all of the variables mentioned above and to study their relationships (see Materials and Methods). Lifestyle was the factor that best explained the distribution of HR in different species (27% of the variance; $P < 0.05$; linear model). Linear models with two variables showed that lifestyle in combination with other variables such as competence capabilities or the fraction of HGT (which alone explains 8% of the variance) can explain over 30% of the HR variation. We then carried out path analysis, which enables the investigation of direct and indirect interactions among variables. The resulting model (Fig. 5) confirmed that lifestyle had the largest direct effect on HR rates, together with variables related to motility and barriers. In addition, this model indicates that, as discussed above, phylogeny did not significantly affect HR levels in a direct way. However, this variable showed an indirect effect mediated through its influence on the genomic characteristics of the different species. Overall, ecological strategy and competence ability combined explained about 32% of the variance in HR events ($P < 0.05$, $r^2 = 0.317$ [analysis of covariance {ANCOVA}, linear model]).

Hot spots of gene exchange and adaptive implications. Genes under positive selection or involved in interspecies genetic transfer have been found to be involved in adaptation to the environment (18, 57, 58). So far, however, we do not understand the possible functional implications of HR at the intraspecific level (16). To uncover prevalent genetic flows and their ecological significance in adaptation or HR processes, we explored the gene content and functional annotations that are overrepresented and underrepresented in HR events across species and lifestyles (Materials and Methods). These analyses revealed patterns (Fig. 6; see also Table S1E) which we explored through

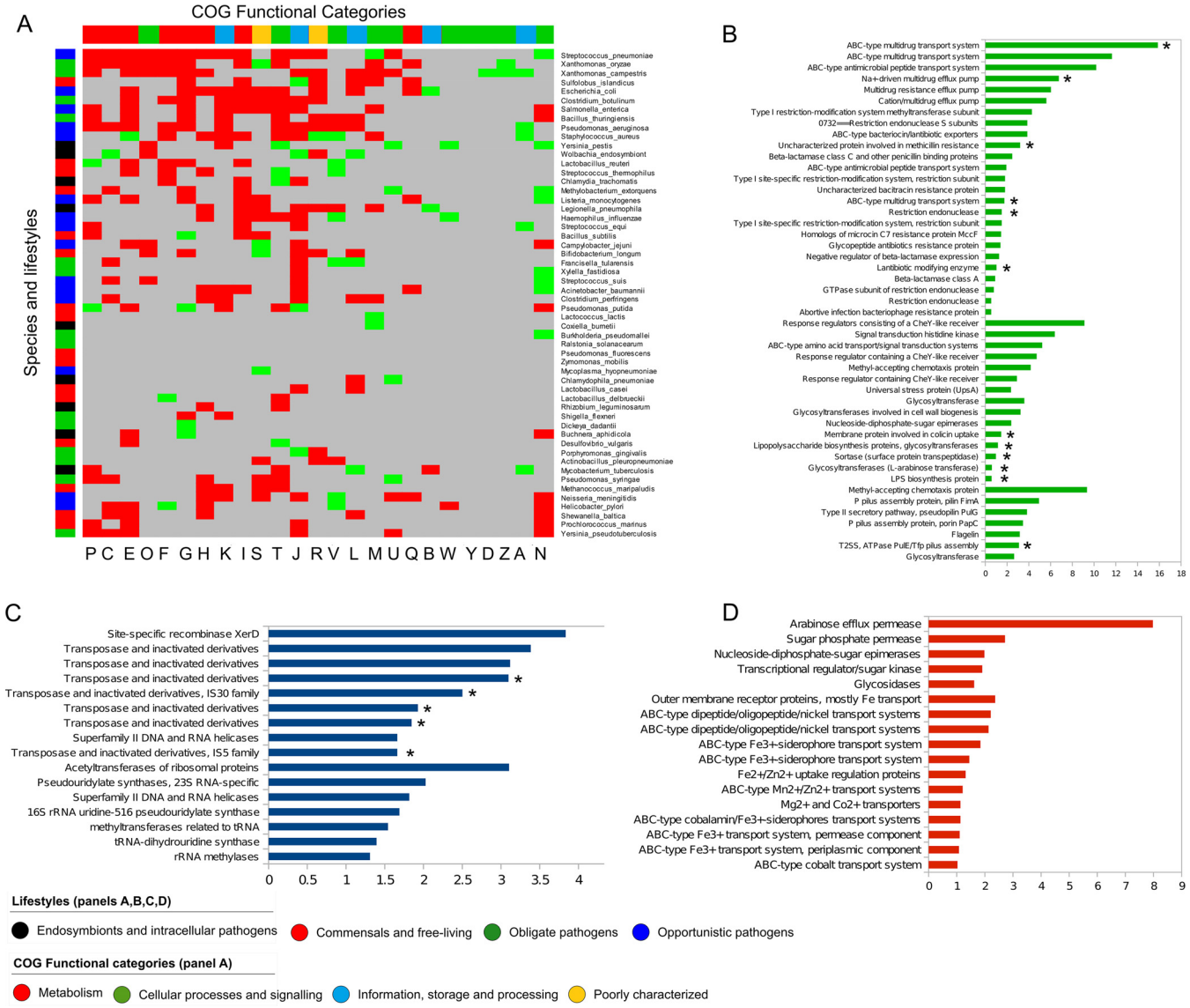

FIG 6 Gene flow and adaptive implications. (A) Heat map for similarity matrix representing the grouping of 54 species analyzed based on their profile and on those with significantly enriched (green) or underrepresented (red) genes ($P < 0.05$ [Fisher's test; FDR correction, <0.1]). The *x* axis shows the layout of the functional categories and the *y* axis the species analyzed. (B to D) Distribution (%) of the most abundant GO terms among HR events and associated with (B) "Information, processing and storage and cellular processes" COG categories (red), (C) "Cellular processes and signaling" COG categories (dark blue), and (D) "Metabolism" COG categories (red). GO terms that presented significant enrichment or underrepresentation are marked with an asterisk (*) ($P < 0.05$, pFDR < 0.05 [Fisher's exact test]).

heat map and clustering analyses (Fig. 6). The identified clusters partially grouped some obligate and opportunistic pathogens as having similar functional patterns of exchanged genes as well as phylogenetically related species that share a lifestyle and have similar HR levels. Overall, the functional category that is most prevalent among exchanged genomic regions is "defense mechanisms" (Clusters of Orthologous Groups [COG] V), which was significantly enriched across all lifestyles. Genes in the "cell motility and secretion" category (COG N) were enriched among obligate and opportunistic pathogens, whereas the latter lifestyle also showed enrichment in "RNA processing and modification" (COG A) ($P < 0.05$ [false-discovery rate {FDR}, <10% [Fisher's test]) (Fig. 6). Notably, the majority of categories related to metabolic functions were underrepresented (Fig. 6). This is in stark contrast with what has been observed in interspecific exchanges, mostly mediated by illegitimate recombination, where metabolism always appears overrepresented (35). To obtain a more detailed insight into the possible

functional implications of HR, we zoomed into the different categories and specifically looked at the lists of enriched terms and genes (Table S1E) in the other two main COG functional categories: "Information, processing and storage" and "cellular processes."

Inspecting the "Information, processing and storage" COG categories, we identified a high abundance of terms connected with HR mechanisms and specific adaptive processes. We observed an overrepresentation among HR regions of the terms "serin recombinase XerD" (4% of enriched terms, the most abundant within enriched terms in the category COG L), and "transposable elements" (25%) (Fig. 6). These enrichments support the idea of a role of transposable elements and *XerD* in HR and integrative processes directly contributing to their own mobilization and likely that of hitchhiking neighboring genes (3, 30, 59). We observed reduced representation of genes related to transcription and protein translation and ribosomal structure (COG K and COG J), mainly in pathogens, while enrichments occurred in terms related to RNA processing (COG A) (Fig. 6). This is in agreement with predictions from the complexity hypothesis (60, 61), which states that extensive horizontal transfer mainly affects operational genes (those that are involved in housekeeping activities and are typically members of large, complex systems) whereas informational genes (those involved in transcription, translation, and related processes) are seldom horizontally transferred. Genes involved in HR within the COG K and COG J categories (8% of the total) were related to cell motility and chemotactic processes as well as encoding methylases, uridylases or rRNA and ribosomal protein acetylases. Of note, 2.5% of the genes associated with HR events were involved in rRNA and tRNA synthesis, which shows that the transfer of informational genes is not necessarily deleterious (6, 62–65). With respect to the enrichment of methylases, methylation of rRNAs has been linked to the acquisition of antibiotic resistance (66–68).

With respect to "cellular processes," we noted that the most abundant or enriched terms were often directly related to resistance, pathogenicity, and adaptive mechanisms. For instance, we found significant enrichments ($P < 0.05$, positive false-discovery rate [pFDR] < 0.05 [Fisher's exact test {FT}]) in multidrug or antibiotic resistance, transport systems, or beta-lactamic resistance, which also plays an important role in cellular communication processes or competition between strains of the same species (69–71). More precisely, we identified significant enrichments in type I RM systems, glycosyltransferases, and O-antigen clusters and virulence factors such as Fe transport systems or type II and IV transport systems. All of them are usually incorporated in the accessory genome (plasmids and GIs) and positively selected (72–76) and move from there to more stable genomic regions (75). Type II and IV transport systems are related to HR events in pathogens (77, 78); virulence capabilities and transferences mediated by transposons (79); and adaptation processes such as virus-host interactions, biofilm formation, secretion of virulence factors, adhesion to surfaces, and DNA transfer through conjugation (80). We found several enriched terms ($P < 0.05$, pFDR < 0.05 [Fisher's exact test]), including glycosyltransferases and epimerases (involved in the synthesis and modification of surface elements [lipopolysaccharides, glycoproteins, and glycolipids]), O-antigen ligases, and lipopolysaccharide biosynthesis, corresponding to gene products that are usually encoded within O-antigen clusters present in GIs of free-living and pathogenic prokaryotes (45, 73, 81, 82). The role of O-antigen diversification in response to environmental viral pressure has been described in free-living organisms (83) as an adaptive mechanism of pathogens to evade immune system (84, 85) or mediating the intraspecific interaction responses in core bank species (71). As for the enrichment in Fe transport systems, most of Fe transporters contribute to environmental adaptation under iron-limiting conditions in both free-living and pathogenic human bacteria (18, 59, 81). Among terms related to "cell process and signaling," the main terms inside "signal transduction" (COG T) involved chemotactic processes, regulation, and universal stress of union gene (UpsA). Taken together, the results of functional analysis of genes enriched in HR regions uncovered enrichments of functions related to important known adaptive pressures and point to a role of selection in shaping patterns of HR.

DISCUSSION

Our comprehensive analysis covering 54 diverse species and using a common methodological framework provides an overall view of the patterns of intraspecific HR in prokaryotes and does so at unprecedented levels of scale and resolution. The validity of our approach is underscored by the congruence of our results with some smaller scale analyses focused on some of the species included in our analysis. However, besides corroborating earlier observations, the broader scope of our approach allows us to unveil global patterns and test trends predicted by different theoretical proposals. In addition, we provide a general model of the interrelationships of the different genomic and ecological factors studied. Overall, our analysis reveals that intraspecific HR is pervasive and occurs in all analyzed species with various levels of impact. After correction for genome size and number of strains was performed, the data revealed that the fraction of the genome affected by HR ranges from 0.5% and 4.13 events/strain (in *Yersinia pestis*) to more than 33.74% and 152.86 events/strain (in *Streptococcus pneumoniae*).

We analyzed a range of genomic and ecological factors that may influence these differences. Among all the factors analyzed, lifestyle emerges as the most relevant, accounting for up to 27% of the variability in HR levels. This explains earlier observations of important differences in HR rates between species of the same genus with different lifestyles (55) and even among lineages of the same species (16, 86). Differences in HR levels between obligate and opportunistic pathogens may relate to differences in the degree of host specialization and in environmental conditions, which is more variable in opportunistic pathogens (13, 42). For instance, we detected the largest HR levels in species such as *S. pneumoniae* and *Helicobacter pylori*, for which our study included strains involved in the same polyclonal infection processes (31, 87). These high levels of HR in pathogens may relate to the selection pressure imposed by the host's immune system to increase the antigenic variability, which has been shown to activate recombination systems and induction of competition and repair systems (44, 75). In this sense, and in addition to having a role in maintaining the cohesion of population clusters, HR can participate in adaptive processes as indicated by the functional analysis of exchanged genes, highlighting the presence of enriched terms directly related to resistance and pathogenicity mechanisms and adaptive processes.

Genomic factors explained a significant part of the observed variability and revealed two clearly differentiated patterns. At one extreme, promiscuous species with large genome sizes and accessory genomes presented higher HR rates and high competence capabilities and presented stronger regulatory mechanisms to control genetic exchange (RM systems and CRISPR-Cas). These promiscuous species tended to be associated with free-living or opportunistic pathogenic lifestyles. At the opposite extreme, we found less-promiscuous species with reduced genomes and low rates of HR and with opposite trends with respect to the aforementioned genetic elements. These species were typically endosymbionts, intracellular pathogens, or obligate pathogens.

Our findings support the idea of a role of competence mechanisms in the exchange of smaller HR events. In contrast, and in accordance with previous suggestions based on individual observations (3, 16, 32) the presence of large recombination events, although uncommon, tends to be associated with the presence of conjugation mechanisms and positive selection. Competence has been considered an adaptive strategy that accelerates adaptation of natural transformants (44, 48, 88). Importantly, we observed that an increase in CRISPR-Cas system type content correlated with a higher proportion of short event fragments among the competent species, supporting the idea of coevolution of the two mechanisms. Finally, we note a general correlation between *rec* gene content and RM systems such as has been previously observed in a limited number of promiscuous species (19, 46, 47) in which the entry of heterologous genetic material and the increase in genome size justify greater control and selection, in accordance with the proteomics constriction theory (38, 89). The greater level of diversity was found in the type III MR system prevalent in those groups (see Fig. S3 in

the supplemental material), which showed higher rates of HR and genome accessory content (Fig. 2 and 3).

Taken together, our analyses provide an overall picture of the different ecological, evolutionary, and genomic factors that modulate the impact of bacterial HR, clarifying the relative importance of each. Although some of the correlations found could be considered to have been expected because they are compatible with current biological knowledge, this should not blur the importance of the finding. First, we have tested many different factors, all of which have been previously proposed to be relevant for HR in the light of current biological knowledge and thus could have been expected to show positive correlations *a priori*. However, our data have not shown support for all of them and, among those that our data have supported, the relative weights are different. For instance, our results support the idea of a stronger role of lifestyle than of phylogenetic background in determining the extent of recombination. The opposite finding—that recombination rates are more similar within a genus or clade regardless of lifestyle—would also make sense from a biological standpoint. Such a finding would simply suggest that recombination rates are more constrained evolutionarily. Thus, our findings are important to define how different factors and processes are weighed in modulating recombination.

Finally, considering their strong relationships with the analyzed ecological, functional, and genomic variables, our data support the idea that HR is one of the main evolutionary mechanisms shaping prokaryotic core genomes. In this regard, our study uncovered many functional links between exchanged genes and specific ecological adaptations, providing the basis of further research. For instance, we observed an enrichment of O-antigen genes among recombined regions, and the diversity of O-antigens is known to determine susceptibility to bacteriophages (90). It could be hypothesized from this that recombination involving these genes may favor adaptation to pressure exerted by phages. Such a hypothesis could be tested, for instance, by *in vitro* evolution experiments exposing mixed populations of bacterial strains to particular phages, whereby recombination in regions encoding O-antigens would support the idea of a selective advantage for some of the bacterial strains. Construction of a detailed list of testable hypotheses that could be derived from our functional enrichments is beyond the scope of this work. Further scrutiny by experts examining the different proteins and species will certainly lead to additional specific and relevant testable predictions.

As we have seen, HR provides a mechanism through which genes or variants can be exchanged, thus providing a substrate upon which selection can act. At the same time, HR acts as a cohesive force keeping population structures intact. In accordance with this role, and extensively supporting the neutral model (10) mentioned above, the species with over 95% ANIb presented r/m values above 0.25, a value that is considered to represent the minimum threshold at which HR acts as a cohesive force on the emerging population clusters. Although, at first glance, support for the neutral model may seem at odds with the reported footprints of selection, this is not necessarily so. As the authors of the neutral model (10) stated, "the use of neutral models of mutation and drift is not a denial of selection, but a recognition that much observed population genetic structure can be explained in simple terms." Thus, the recombination would occur stochastically and would result in the emergence of population clusters even in the absence of selection. However, due to differences in fitness between the different recombined genomes in a given environment, selection would act to determine which population clusters, and which individuals within each cluster, are more likely to survive.

MATERIALS AND METHODS

Genome data set. Bacterial and archaeal genomes fully assembled into a single superscaffold were downloaded from the NCBI FTP site (https://www.ncbi.nlm.nih.gov/genomes). We considered only species for which three or more fully sequenced genomes of different strains were available. This resulted in a total of 338 genomes (325 bacterial genomes and 13 archaeal genomes) from 54 species, each including 3 to 15 genomes (Fig. 1). Additionally, 19 genomes from strains belonging to the same genus

but not the same species were selected and included in five control groups (16S ANI, <98.7%) (22) that were used to estimate false-positive rates in HR detection methods (see below).

We validated the fitting of each clade of selected strains to the 16S divergence criterion to define species using rRNA 16S pairwise alignments using SILVA Incremental Aligner (SINA) 1.2.11v (91), ensuring that the average nucleotide identity (ANI) within the 16S sequence was within the species range (similarity of >98.7%) (21) or, in the case of the controls, within the genus range only (similarity of <98.7%) (21). Species were classified into the following four main groups of lifestyles, as defined previously (13): (i) intracellular pathogens and symbionts, (ii) opportunistic pathogens, (iii) obligate pathogens, and (iv) free-living microbes.

Whole-genome alignments: positional orthologs and values corresponding to genome-wide average nucleotide identity based on BLAST (ANIb). The genomes of the species were aligned using the progressive function from the Mauve Package 2.3.1v using default parameters (92). The output consisted of the alignment map showing the local colinear blocks (LBCs) and the main rearrangements and of the sequences contained in the aligned blocks (eXtended Multi-FastA [*xmfa] format) and the positional orthologs therein (*ort files). The output files were parsed with a python script which selected the core genome (genes shared among all the strains) and accessory genomes (genes missing in some strains) (93, 94).

Genome-wide average nucleotide identity based on BLAST (ANIb) between strains within species. The genomic sequences within each species were compared in a pairwise mode using the Nucmer function from the MUMmer package v3.0 (95). The resulting coordinate files (*coords files) were parsed with a python script to calculate the percentage of identity for the aligned sequences. For this, fragments that overlapped over 10% of their length were joined, and those contained within a larger fragment were filtered out. An overall nucleotide identity value based on BLAST (ANIb) was obtained for each comparison as the average of each aligned fragment weighted by its size.

Detection and characterization of recombination events. RDP4 v4.15 (24) was used to detect and characterize recombination events. This program implements different recombination detection methods to both detect and characterize the recombination events that are evident within a sequence alignment without any prior user indication of a nonrecombinant set of reference sequences. The implemented algorithms use a combination of methods based on partitioning schemes, dynamic scanning window strategy-based algorithms, and testing schemes that generally consist of two steps of analysis, the first performed to detect changes in the phylogenetic sequence relationships between partitions and the second to statistically test the approximate significance of these changes. Genomic alignments obtained via the *xmfa format described above were used as input files. Recombination events were predicted in a two-step procedure. In the first, exploratory phase, the following four different methods were run: RDP (96), GENECONV (97), MaxCHI (98), and Chimera (99). In the second phase, the program rescanned every detected event more thoroughly with the RDP, GENECONV, MaxCHI, Chimera, and 3Seq algorithms (100). The window size settings for the RDP algorithm was adjusted to 90 nucleotides, and the number of variable sites per window was set to 210 in the case of MaxCHI. For the remaining methods, default parameters were applied. In all cases, a cutoff probability (*P*) value of <0.001 was used. Only recombination signals detected by at least three of the five methods were considered. Finally, the predicted recombination events were manually curated, and the breakpoints were inferred using the MaxCHI method (which is considered the most accurate breakpoint detection method among the five nonparametric methods implemented in RDP3) (24). The following variables were recorded for each comparison: fraction of genome recombined, number of recombination events per strain, and distribution of the sizes of the recombination events. To obtain a rarefaction curve in terms of the number of HR events detected depending on the number of compared genomes, we repeated the analysis with subsets of 10 randomly selected *E. coli* strains from our set.

Experimental design and variables. We evaluated several genomic, evolutionary, and ecological variables obtained from different databases (see Table S1E in the supplemental material). Genomic variables included the following: (i) size of core genome (number of kilobases and number of genes); (ii) size of GIs; (iii) number of ribosomal operons; (iv) tRNA content; (v) presence of elements that favor intraspecific DNA exchange (also known as motility variables) such as content of RM systems (type I, II, II, and IV) and competence capabilities (*com* gene content); and (vi) presence of elements that make interspecific DNA exchange difficult (i.e. barriers) such as CRISPR-Cas system content. Evolutionary variables included the following: (i) phylogenetic position; (ii) ANIb; (iii) ratio of nonsynonymous substitutions to synonymous substitutions (dN/dS); (iv) number of recombinant events by strain (events/strain); (v) recombination/mutation ratio (r/m); and (vi) proportion of recombined genome. The relevant qualitative ecological variables included the following: competence capabilities and lifestyle classification from previous MLSA studies (13, 16) and JGI database metadata assigning each set of strains to one of the four main groups mentioned above.

Functional analysis. Functional annotations were retrieved for all 338 strains from the Integrate Microbial Genomes (IMG) database (Joint Genomics Institute) (101). Sequences of the predicted recombined fragments were retrieved from the genome FASTA files using the positional coordinates provided by RDP4 v4.15 (24). For those sequences, *de novo* gene prediction was performed employing the Integrative Microbial Genomes (IMG) system (Joint Genomics Institute) (101). Recombinant regions shorter than 10 kb were annotated with an algorithm implemented for metagenomic data (IMG/M ER), and a genomic algorithm (IMG ER) (101) was used for those longer than 10 kb. Functional terms from the Clusters of Orthologous Groups (COG) (102) and Gene Ontology (GO) (103) databases were retrieved for both genome and recombination fragments using the JGI platform and Blast2GO (B2G) software (104,

105), respectively. In the latter case, hits with an E value lower than 1×10^{-20} and amino acid sequence identity higher than 55% were considered. Finally, KEGG EC numbers (106) were retrieved.

Statistical analyses. Tests for functional enrichment of genes contained in recombination fragments versus the genomic background were performed using Fisher's exact test (FT) with the COG and GO annotation terms. The Fisher's test was performed for each of the functional categories in each species by applying a false-discovery rate (FDR) correction and designating 0.05 the *P* value threshold for over- or underrepresentations, as implemented in the Gossip tool from Blast2Go. The distributions of the variables described above were compared using Kruskal-Wallis/Jonckheere-Terpstra tests as implemented in the SPSSv22 statistical package complemented by multiple comparisons or Bonferroni adjustment.

Bivariate and partial correlations were used to explore relationships between quantitative variables using the SPSSv22 statistical package. For this purpose, parametric and nonparametric tests were run using Pearson and Spearman correlation coefficients, respectively, to assess bilateral significance (marginally significant, <0.1; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). Mantel-Haenszel and partial Mantel-Haenszel tests were applied to Euclidean distance arrays generated from related recombinant variables (events/strain and proportion of genome recombined), motility variables (competence and RM content), barrier variables (CRISPR-Cas and HGT), and phylogeny genomic and lifestyle class. *P* values of less than 0.05 were considered significant. For those factors involved in the distribution of the variable homologous recombination data, a general model was built from the results of the Mantel-Haenszel test by means of a path analysis.

Evolutionary genomics analysis. We reconstructed the clonal genealogy using ClonalFrame 1.2v (27) and the genomic alignments sorted with MAUVE 2.3.1v as input. Core alignments were extracted by keeping only those regions that were aligned for all genomes over at least 500 bp. Three independent ClonalFrame runs were performed, each consisting of 40,000 iterations. The first half of these iterations was discarded as representative of Markov chain Monte Carlo (MCMC) burn-in. The convergence of the tree runs was checked by manual comparison, making sure that they produced consistent estimates of the clonal genealogy and of the global parameters r/m (where r = rate of recombination and m = rate of mutation) (13) and rho/theta (σ/θ) (107) (where σ and θ are the rates of occurrence of recombination and mutation, respectively) with 95% credibility. The σ/θ ratio is a measure of the frequency at which recombination occurs relative to mutation, and the r/m ratio is a measure of the rates at which nucleotides become substituted as a result of recombination or mutation and estimates the relative effects of HR on genetic diversification of populations.

The amino acid sequences for the positional orthologous genes were retrieved from the GenBank files. Pairwise sequence alignments between selected strains of each species were performed with MUSCLE 3.8v (108). The resulting alignments were reverse translated to codon-based nucleotide alignments using trimAL v1.3 (109) and the corresponding coding sequences. Finally, dN/dS values were obtained using the CodeML function (pairwise mode with model 1 nonsynonymous [NS] sites [0 parameters]) of PAML package 4.4v (110).

Ethics approval and consent to participate. No data from humans were used in the work described in this article.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02494-18.

FIG S1, TIF file, 0.6 MB.
FIG S2, TIF file, 0.4 MB.
FIG S3, TIF file, 0.3 MB.
FIG S4, TIF file, 0.3 MB.
FIG S5, TIF file, 0.4 MB.
TABLE S1, XLS file, 0.4 MB.

ACKNOWLEDGMENTS

The group of J.A. is funded by grant CLG2015_66686-C3-03 from the Spanish Ministry of Economy and Competitiveness (MINECO), which is cofinanced with FEDER support from the European Union. P.G.-T. was an FPI-MINECO fellow associated with project CGL2012-39627-C03-01 (to J.A.). The T.G. group acknowledges support from the Spanish Ministry of Economy and Competitiveness (grants "Centro de Excelencia Severo Ochoa 2013-2017" SEV-2012-0208 and BFU2015-67107; cofounded by European Regional Development Fund [ERDF]); from the CERCA Program/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) (grant SGR857); and from the European Union's Horizon 2020 research and innovation program under grant agreement ERC-2016-724173 (Marie Sklodowska-Curie grant agreement no. H2020-MSCA-ITN-2014-642095).

P.G.-T. and F.R.-M. performed computational experiments. P.G.-T. T.G., and J.A. wrote

the paper. T.G. and J.A. supervised the research. All of us read and approved the final manuscript.

We declare that we have no competing interests.

REFERENCES

1. Mira A, Klasson L, Andersson SGE. 2002. Microbial genome evolution: sources of variability. Curr Opin Microbiol 5:506–512. https://doi.org/10.1016/S1369-5274(02)00358-2.
2. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. Proc Natl Acad Sci U S A 102:13950–13955. https://doi.org/10.1073/pnas.0506758102.
3. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721. https://doi.org/10.1038/nrmicro1234.
4. Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29:170–175. https://doi.org/10.1016/j.tig.2012.12.006.
5. Konstantinidis KT, DeLong EF. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J 2:1052–1065. https://doi.org/10.1038/ismej.2008.62.
6. Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF. 2007. Searching for species in haloarchaea. Proc Natl Acad Sci U S A 104:14092–14097. https://doi.org/10.1073/pnas.0706358104.
7. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65. https://doi.org/10.1038/nature08821.
8. Bhaya D, Grossman AR, Steunou A-S, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF. 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J 1:703–713. https://doi.org/10.1038/ismej.2007.46.
9. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. https://doi.org/10.1038/nature02340.
10. Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480. https://doi.org/10.1126/science.1127573.
11. Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos Trans R Soc Lond B Biol Sci 361:1985–1996. https://doi.org/10.1098/rstb.2006.1918.
12. Cohan FM. 2002. Sexual isolation and speciation in bacteria. Genetica 116:359–370. https://doi.org/10.1023/A:1021232409545.
13. Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208. https://doi.org/10.1038/ismej.2008.93.
14. Caro-Quintero A, Konstantinidis KT. 2012. Bacterial species may exist, metagenomics reveal. Environ Microbiol 14:347–355. https://doi.org/10.1111/j.1462-2920.2011.02668.x.
15. Caro-Quintero A, Deng J, Auchtung J, Brettar I, Höfle MG, Klappenbach J, Konstantinidis KT. 2011. Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea. ISME J 5:131–140. https://doi.org/10.1038/ismej.2010.93.
16. Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322. https://doi.org/10.1016/j.tim.2010.04.002.
17. Martin DP, Lemey P, Posada D. 2011. Analysing recombination in nucleotide sequences. Mol Ecol Resour 11:943–955. https://doi.org/10.1111/j.1755-0998.2011.03026.x.
18. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, Falush D. 2016. The landscape of realized homologous recombination in pathogenic bacteria. Mol Biol Evol 33:456–471. https://doi.org/10.1093/molbev/msv237.
19. Oliveira PH, Touchon M, Rocha EPC. 2016. Regulation of genetic flux between bacteria by restriction–modification systems. Proc Natl Acad Sci U S A 113:5658–5663. https://doi.org/10.1073/pnas.1603257113.
20. Bobay L-M, Ochman H. 2017. Impact of recombination on the base composition of bacteria and archaea. Mol Biol Evol 34:2627–2636. https://doi.org/10.1093/molbev/msx189.
21. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645. https://doi.org/10.1038/nrmicro3330.
22. Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440. https://doi.org/10.1038/nrmicro1872.
23. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol 1:vev003. https://doi.org/10.1093/ve/vev003.
24. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462–2463. https://doi.org/10.1093/bioinformatics/btq467.
25. Martin DP, Murrell B, Khoosal A, Muhire B. 2017. Detecting and analyzing genetic recombination using RDP4. Methods Mol Biol 1525:433–460. https://doi.org/10.1007/978-1-4939-6622-6_17.
26. Didelot X, Wilson DJ, Bryant D, Quail M, Cockfield J. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041.
27. Didelot X, Falush D. 2006. Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266. https://doi.org/10.1534/genetics.106.063305.
28. Hanage WP, Fraser C, Spratt BG. 2006. The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol 239:210–219. https://doi.org/10.1016/j.jtbi.2005.08.035.
29. Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 6:97–112. https://doi.org/10.1016/j.meegid.2005.02.003.
30. Mau B, Glasner JD, Darling AE, Perna NT. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol 7:R44. https://doi.org/10.1186/gb-2006-7-5-r44.
31. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, Barbadora K, Klimke W, Dernovoy D, Tatusova T, Parkhill J, Bentley SD, Post JC, Ehrlich GD, Hu FZ. 2007. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189:8186–8195. https://doi.org/10.1128/JB.00690-07.
32. Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7:e1002129. https://doi.org/10.1371/journal.ppat.1002129.
33. Polz MF, Hanage WP. 2013. Quantitative and theoretical microbial population biology, p 31–42. *In* Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed), The prokaryotes: prokaryotic communities and ecophysiology. Springer, New York, NY. https://doi.org/10.1007/978-3-642-30123-0.
34. Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene

transfer frequency at different phylogenetic depths. Mol Biol Evol 28:1057–1074. https://doi.org/10.1093/molbev/msq297.
35. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609. https://doi.org/10.1101/gr.115592.110.
36. Makarova KS, Wolf YI, Koonin EV. 2013. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res 41:4360–4377. https://doi.org/10.1093/nar/gkt157.
37. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60:174–182. https://doi.org/10.1007/s00239-004-0046-3.
38. Garcia-Gonzalez A, Vicens L, Alicea M, Massey SE. 2013. The distribution of recombination repair genes is linked to information content in bacteria. Gene 528:295–303. https://doi.org/10.1016/j.gene.2013.05.082.
39. El Karoui M, Biaudet V, Schbath S, Gruss A. 1999. Characteristics of Chi distribution on different bacterial genomes. Res Microbiol 150:579–587. https://doi.org/10.1016/S0923-2508(99)00132-1.
40. Wigley DB. 2013. Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nat Rev Microbiol 11:9–13. https://doi.org/10.1038/nrmicro2917.
41. Krajewski WW, Fu X, Wilkinson M, Cronin NB, Dillingham MS, Wigley DB. 2014. Structural basis for translocation by AddAB helicase–nuclease and its arrest at χ sites. Nature 508:416–419. https://doi.org/10.1038/nature13037.
42. Michod RE, Bernstein H, Nedelcu AM. 2008. Adaptive value of sex in microbial pathogens. Infect Genet Evol 8:267–285. https://doi.org/10.1016/j.meegid.2008.01.002.
43. Rocha EPC, Cornet E, Michel B. 2005. Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet 1:e15. https://doi.org/10.1371/journal.pgen.0010015.
44. Claverys J-P, Prudhomme M, Martin B. 2006. Induction of competence regulons as a general response to stress in Gram-positive bacteria. Annu Rev Microbiol 60:451–475. https://doi.org/10.1146/annurev.micro.60.080805.142139.
45. Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2:414–424. https://doi.org/10.1038/nrmicro884.
46. Vasu K, Nagaraja V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev 77:53–72. https://doi.org/10.1128/MMBR.00044-12.
47. Corvaglia AR, Francois P, Hernandez D, Perron K, Linder P, Schrenzel J. 2010. A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. Proc Natl Acad Sci 107:11954–11958. https://doi.org/10.1073/pnas.1000489107.
48. Jorth P, Whiteley M. 2012. An evolutionary link between natural transformation and CRISPR. mBio 3:e00309-12. https://doi.org/10.1128/mBio.00309-12.
49. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. Science 311:1768–1770. https://doi.org/10.1126/science.1122050.
50. Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F. 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J 1:235–245. https://doi.org/10.1038/ismej.2007.35.
51. Tully BJ, Emerson JB, Andrade K, Brocks JJ, Allen EE, Banfield JF, Heidelberg KB. 2015. De novo sequences of Haloquadratum walsbyi from Lake Tyrrell, Australia, reveal a variable genomic landscape. Archaea 2015:875784. https://doi.org/10.1155/2015/875784.
52. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci 102:2567–2572. https://doi.org/10.1073/pnas.0409727102.
53. Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philos Trans R Soc B Biol Sci 361:1929–1940. https://doi.org/10.1098/rstb.2006.1920.
54. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Microevolutionary genomics of bacteria. Theor Popul Biol 61:435–447. https://doi.org/10.1006/tpbi.2002.1588.
55. Larsson P, Elfsmark D, Svensson K, Wikström P, Forsman M, Brettin T, Keim P, Johansson A. 2009. Molecular evolutionary consequences of niche restriction in Francisella tularensis, a facultative intracellular pathogen. PLoS Pathog 5:e1000472. https://doi.org/10.1371/journal.ppat.1000472.
56. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239:226–235. https://doi.org/10.1016/j.jtbi.2005.08.037.
57. Orsi RH, Sun Q, Wiedmann M. 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes. BMC Evol Biol 8:233. https://doi.org/10.1186/1471-2148-8-233.
58. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. 2011. Interplay of recombination and selection in the genomes of Chlamydia trachomatis. Biol Direct 6:28. https://doi.org/10.1186/1745-6150-6-28.
59. Hacker J, Carniel E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep 2:376–381. https://doi.org/10.1093/embo-reports/kve097.
60. Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801–3806. https://doi.org/10.1073/pnas.96.7.3801.
61. Soucy SM, Huang J, Peter Gogarten J. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962.
62. Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle WF. 2007. Intragenomic heterogeneity and intergenomic recombination among Vibrio parahaemolyticus 16S rRNA genes. Microbiology 153:2640–2647. https://doi.org/10.1099/mic.0.2007/009175-0.
63. Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. 2009. Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus. Genome Biol Evol 1:325–339. https://doi.org/10.1093/gbe/evp032.
64. Williams D, Gogarten JP, Papke RT. 2012. Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol Evol 4:1223–1244. https://doi.org/10.1093/gbe/evs098.
65. Cheng K, Rong X, Huang Y. 2016. Widespread interspecies homologous recombination reveals reticulate evolution within the genus Streptomyces. Mol Phylogenet Evol 102:246–254. https://doi.org/10.1016/j.ympev.2016.06.004.
66. Benítez-Páez A, Cárdenas-Brito S, Corredor M, Villarroya M, Armengod ME. 2013. Impairing methylations at ribosome RNA, a point mutation-dependent strategy for aminoglycoside resistance: the rsmG case. Biomédica 34:41–49. https://doi.org/10.1590/S0120-41572014000500006.
67. Kotra LP, Haddad J, Mobashery S. 2000. Aminoglycosides: perspectives on mechanisms of action and resistance and strategies to counter resistance. Antimicrob Agents Chemother 44:3249–3256. https://doi.org/10.1128/AAC.44.12.3249-3256.2000.
68. Wu Q, Zhang Y, Han L, Sun J, Ni Y. 2009. Plasmid-mediated 16S rRNA methylases in aminoglycoside-resistant Enterobacteriaceae isolates in Shanghai, China. Antimicrob Agents Chemother 53:271–272. https://doi.org/10.1128/AAC.00748-08.
69. Davies J. 2006. Are antibiotics naturally antibiotics? J Ind Microbiol Biotechnol 33:496–499. https://doi.org/10.1007/s10295-006-0112-5.
70. Bernier SP, Surette MG. 2013. Concentration-dependent activity of antibiotics in natural environments. Front Microbiol 4:20. https://doi.org/10.3389/fmicb.2013.00020.
71. González-Torres P, Pryszcz LP, Santos F, Martínez-García M, Gabaldón T, Antón J. 2015. Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing. Appl Environ Microbiol 81:8445–8456. https://doi.org/10.1128/AEM.02690-15.
72. Bellanger X, Payot S, Leblond-Bourget N, Guédon G. 2014. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol Rev 38:720–760. https://doi.org/10.1111/1574-6976.12058.
73. Fernández-Gómez B, Fernàndez-Guerra A, Casamayor EO, González JM, Pedrós-Alió C, Acinas SG. 2012. Patterns and architecture of genomic islands in marine bacteria. BMC Genomics 13:347. https://doi.org/10.1186/1471-2164-13-347.
74. Furuta Y, Abe K, Kobayashi I. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. Nucleic Acids Res 38:2428–2443. https://doi.org/10.1093/nar/gkp1226.
75. Makarova KS, Wolf YI, Snir S, Koonin EV. 2011. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol 193:6039–6056. https://doi.org/10.1128/JB.05535-11.

76. Takahashi N, Ohashi S, Sadykov MR, Mizutani-Ui Y, Kobayashi I. 2011. IS-linked movement of a restriction-modification system. PLoS One 6:e16554. https://doi.org/10.1371/journal.pone.0016554.
77. Guy L, Nystedt B, Sun Y, Näslund K, Berglund EC, Andersson SGE. 2012. A genome-wide study of recombination rate variation in Bartonella henselae. BMC Evol Biol 12:65. https://doi.org/10.1186/1471-2148-12-65.
78. Nandi T, Holden MTG, Didelot X, Mehershahi K, Boddey JA, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. 2015. Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. Genome Res 25:129–141. https://doi.org/10.1101/gr.177543.114.
79. Ambur OH, Davidsen T, Frye SA, Balasingham SV, Lagesen K, Rognes T, Tønjum T. 2009. Genome dynamics in major bacterial pathogens. FEMS Microbiol Rev 33:453–470. https://doi.org/10.1111/j.1574-6976.2009.00173.x.
80. Melville S, Craig L. 2013. Type IV pili in Gram-positive bacteria. Microbiol Mol Biol Rev 77:323–341. https://doi.org/10.1128/MMBR.00063-12.
81. Gonzaga A, Martin-Cuadrado AB, López-Pérez M, Mizuno CM, García-Heredia I, Kimes NE, Lopez-García P, Moreira D, Ussery D, Zaballos M, Ghai R, Rodriguez-Valera F. 2012. Polyclonality of concurrent natural populations of Alteromonas macleodii. Genome Biol Evol 4:1360–1374. https://doi.org/10.1093/gbe/evs112.
82. Rodriguez-Valera F, Martin-Cuadrado A-B, López-Pérez M. 2016. Flexible genomic islands as drivers of genome evolution. Curr Opin Microbiol 31:154–160. https://doi.org/10.1016/j.mib.2016.03.014.
83. López-Pérez M, Gonzaga A, Rodriguez-Valera F. 2013. Genomic diversity of 'deep ecotype' Alteromonas macleodii isolates: evidence for pan-mediterranean clonal frames. Genome Biol Evol 5:1220–1232. https://doi.org/10.1093/gbe/evt089.
84. Schmidt MA, Riley LW, Benz I. 2003. Sweet new world: glycoproteins in bacterial pathogens. Trends Microbiol 11:554–561. https://doi.org/10.1016/j.tim.2003.10.004.
85. Wang X, Quinn PJ. 2010. Lipopolysaccharide: biosynthetic pathway and structure modification. Prog Lipid Res 49:97–107. https://doi.org/10.1016/j.plipres.2009.06.002.
86. den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M. 2010. Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:688. https://doi.org/10.1186/1471-2164-11-688.
87. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S. 2011. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 108:5033–5038. https://doi.org/10.1073/pnas.1018444108.
88. Seitz P, Blokesch M. 2013. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. FEMS Microbiol Rev 37:336–363. https://doi.org/10.1111/j.1574-6976.2012.00353.x.
89. Massey SE. 2013. Proteome size as the major factor determining mutation rates. Proc Natl Acad Sci 110:E858–E859. https://doi.org/10.1073/pnas.1219306110.
90. Villamor J, Ramos-Barbero MD, González-Torres P, Gabaldón T, Rosselló-Móra R, Meseguer I, Martínez-García M, Santos F, Antón J. 2017. Characterization of ecologically diverse viruses infecting co-occurring strains of cosmopolitan hyperhalophilic Bacteroidetes. ISME J https://doi.org/10.1038/ismej.2017.175.
91. Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829. https://doi.org/10.1093/bioinformatics/bts252.
92. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. https://doi.org/10.1371/journal.pone.0011147.
93. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev 15:589–594. https://doi.org/10.1016/j.gde.2005.09.006.
94. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. 2010. The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 13:45–57. https://doi.org/10.2436/20.1501.01.110.
95. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
96. Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562–563. https://doi.org/10.1093/bioinformatics/16.6.562.
97. Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–225. https://doi.org/10.1006/viro.1999.0056.
98. Smith JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol 34:126–129.
99. Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98:13757–13762. https://doi.org/10.1073/pnas.241370698.
100. Boni MF, Posada D, Feldman MW. 2006. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176:1035–1047. https://doi.org/10.1534/genetics.106.068874.
101. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:115–122. https://doi.org/10.1093/nar/gkr1044.
102. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. https://doi.org/10.1093/nar/28.1.33.
103. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556.
104. Conesa A, Götz S. 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:1. https://doi.org/10.1155/2008/619832.
105. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435. https://doi.org/10.1093/nar/gkn176.
106. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34. https://doi.org/10.1093/nar/28.1.27.
107. Milkman R, Bridges MM. 1990. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. Genetics 126:505–517.
108. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
109. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
110. Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088.